# Why multigrid can be effective in optimization

Michael Lewis

College of William & Mary

ICASE, NASA Langley Research Center

Joint work with Stephen G. Nash, George Mason University

This work was supported by the National Aeronautics and Space Administration under NASA Contract NAS1-97046.

## The problem of interest

The nonlinear program (NLP):

$$\begin{array}{ll} \underset{a}{\text{minimize}} & F(a) = f(a, u(a)) \\ \text{subject to} & C(a) = C(a, u(a)) \geq 0 \end{array}$$

where $S(a, u(a) = 0$ is the governing PDE.

Terminology:

design variables: $a \in X$, finite- or $\infty$-dimensional
state variables: $u \in U$, $\infty$-dimensional

The computational cost of optimization is determined by the discretization of the governing differential equations.

A finer discretization means greater accuracy, but more work.

## Another type of NLP related to discretized problems

The NLP:

$$\begin{array}{ll} \underset{u}{\text{minimize}} & f(u) \\ \text{subject to} & c(u) \geq 0 \end{array}$$

**Example.** Minimize the area of a surface (graph) with prescribed boundary.

$$\begin{array}{lll} \underset{u}{\text{minimize}} & \displaystyle\int_\Omega (1 + \| \nabla u \|^2)^{1/2} \, dx & \\ \text{subject to} & u = \phi & \text{on } \partial\Omega \end{array}$$

The stationarity condition is the minimal surface equation.

# This talk

1. The class of nonlinear programs of interest
2. A multigrid method
3. Some model problems and numerical results
4. Why multigrid might work:
   - The nature of the reduced Hessian
5. Interaction with truncated conjugate gradients

We assume here that the design variable $a$ is a discretized quantity $a_h$.

## Optimization of systems governed by differential equations

**General theme:** The governing PDE and the NLP interact in many interesting ways, both analytically and computationally.

**This talk:** An optimization problem may be better suited to a multigrid approach than its governing p.d.e.

- MG/Opt

  *Model problems for the multigrid optimization of systems governed by differential equations*, RML and S. G. Nash, submitted to *SIAM J. on Scientific Computing*.

  *A Multigrid Approach to the Optimization of Systems Governed by Differential Equations*, RML and S. G. Nash, AIAA paper 2000-4890.

- For a related approach, see

  *Optimization with variable-fidelity models applied to wing design*, N. M. Alexandrov, RML, C. R. Gumbert, L. L. Green, P. A. Newman, *J. of Aircraft*, Nov–Dec 2001.

## MG/Opt overview

The MG/Opt multigrid approach to the nonlinear program:

- Multigrid: recursively use coarse grid problems to generate search directions for finer grid problems
- Use a line search on fine grid
- Convergence can be guaranteed
- Inspired by multigrid for elliptic p.d.e. and by globalization techniques in nonlinear programming
- Applicable when $S(a, u) = 0$ is not especially amenable to multigrid (e.g., hyperbolic p.d.e.)
- Optimization problem better suited to multigrid than underlying differential equation

## Multigrid for linear elliptic p.d.e.

For a linear system $Ax = b$

- If on coarsest grid, solve and return
- Apply $k_1$ iterations of an iterative method
- Form residual $r = b - Ax$
- Solve (recursively) coarse-grid version of $Ae = r$, and update solution to fine grid
- Set $x \leftarrow x + e$
- Apply $k_2$ iterations of an iterative method

## Properties of linear multigrid

- Storage: about 2 times the storage of the fine-grid problem $(N + N/2 + \ldots)$
- Computation: 1 MG iteration $=$ about 4 fine-grid iterations
- Convergence: on "appropriate" problems, no. of MG iterations is independent of the fine-grid resolution
- Linear convergence rate

## Multigrid optimization (MG/Opt) algorithm

- Originally developed for unconstrained variational problems

$$\underset{u}{\text{minimize}}\, f_h(u)$$

- Here, the p.d.e. $S_h(a, u) = 0$ is solved for $u_h(a)$ (given $a$)
- In many cases, no. of design variables $a$ is fixed
  - semi-coarsening in states $u_h$ only

  Here we assume $a$ is a discretized quantity $a_h$.
- Motivated by full approximation scheme applied to optimality conditions:

$$\nabla_a f_h(a, u_h(a)) = 0$$

- Alternatively, one can motivate the algorithm via NLP considerations

# MG/Opt algorithm

Notation

- $h$: fine grid mesh, $H$: coarse grid mesh
- $I_h^H$: downdate, $I_H^h$: update, of $a$
- $u_h$: fine-grid vector, $u_H$: coarse-grid vector
- $F_h$: fine-grid objective function

$$F_h(a_h) = f(a_h, u_h(a_h))$$

- $F_H$: coarse-grid objective function
- $g_h(a)$: fine-grid gradient
- $g_H(a)$: coarse-grid gradient
- $g_1 = g(a_1, u_1(a_1))$, etc.

# MG/Opt algorithm

1. If coarsest grid, solve $\underset{a}{\text{minimize}}\ f_h(a, u_h(a))$; else:
2. partially minimize $F_h(a)$ to get $a_1$
3. set $\bar{a}_1 = I_h^H a_1$
4. compute $v = \bar{g}_1 - I_h^H g_1$
5. recursively minimize $F_H(a) - v^T a$ (with initial guess: $\bar{a}_1$, result: $\bar{a}_2$) subject to bound constraints on the solution (used to guarantee convergence)
6. compute $e_2 = I_H^h(\bar{a}_2 - \bar{a}_1)$
7. line search: $a_2 \leftarrow a_1 + \alpha e_2$
8. partially minimize $F_h(a)$ to get $a_3$

## User requirements

- Subroutine to solve $S(a, u) = 0$ for $u$ given $a$
- Subroutine to evaluate $F_h(a)$ and $\nabla_a F_h(a)$ for any grid $h$
- Subroutines to implement downdate $I_H^h$ and update $I_h^H$ operators
  - Should satisfy $I_H^h = \text{const} \times (I_h^H)^T$ (standard)

# $\nabla^2 F(a)$ as a reduced Hessian

Formally, $\nabla^2 F(a)$ is the reduced Hessian associated with the formulation

$$\begin{array}{ll} \underset{(a,u)}{\text{minimize}} & f(a,u) \\ \text{subject to} & S(a,u) = 0 \end{array}$$

Let $W$ be the following basis for the nullspace of the linearized constraints:

$$\left[ \begin{array}{cc} S_a & S_u \end{array} \right] W = \left[ \begin{array}{cc} S_a & S_u \end{array} \right] \left( \begin{array}{c} I \\ -S_u^{-1} S_a \end{array} \right) = 0.$$

Define the Lagrangian $L(a,u;\lambda) = f(a,u) + \langle \lambda,\ S(a,u) \rangle$.

Then

$$\boxed{\nabla^2 F(a) = W^T \left( \nabla^2_{(a,u)} L((a, u(a); \lambda) \right) W.}$$

# $\nabla^2 F(a)$ in detail

Let

$$\nabla^2_{(a,u)} L((a,u;\lambda) = \nabla^2_{(a,u)} f(a,u) + \nabla^2_{(a,u)} S(a,u)\lambda = \begin{pmatrix} L_{aa} & L_{au} \\ L_{ua} & L_{uu} \end{pmatrix}.$$

Then

$$\boxed{\nabla^2 F = S_a^T S_u^{-T} L_{uu} S_u^{-1} S_a + L_{au} S_u^{-1} S_a + S_a^T S_u^{-T} L_{ua} + L_{aa}.}$$

## Model Problem: Dirichlet to Neumann map

Minimize

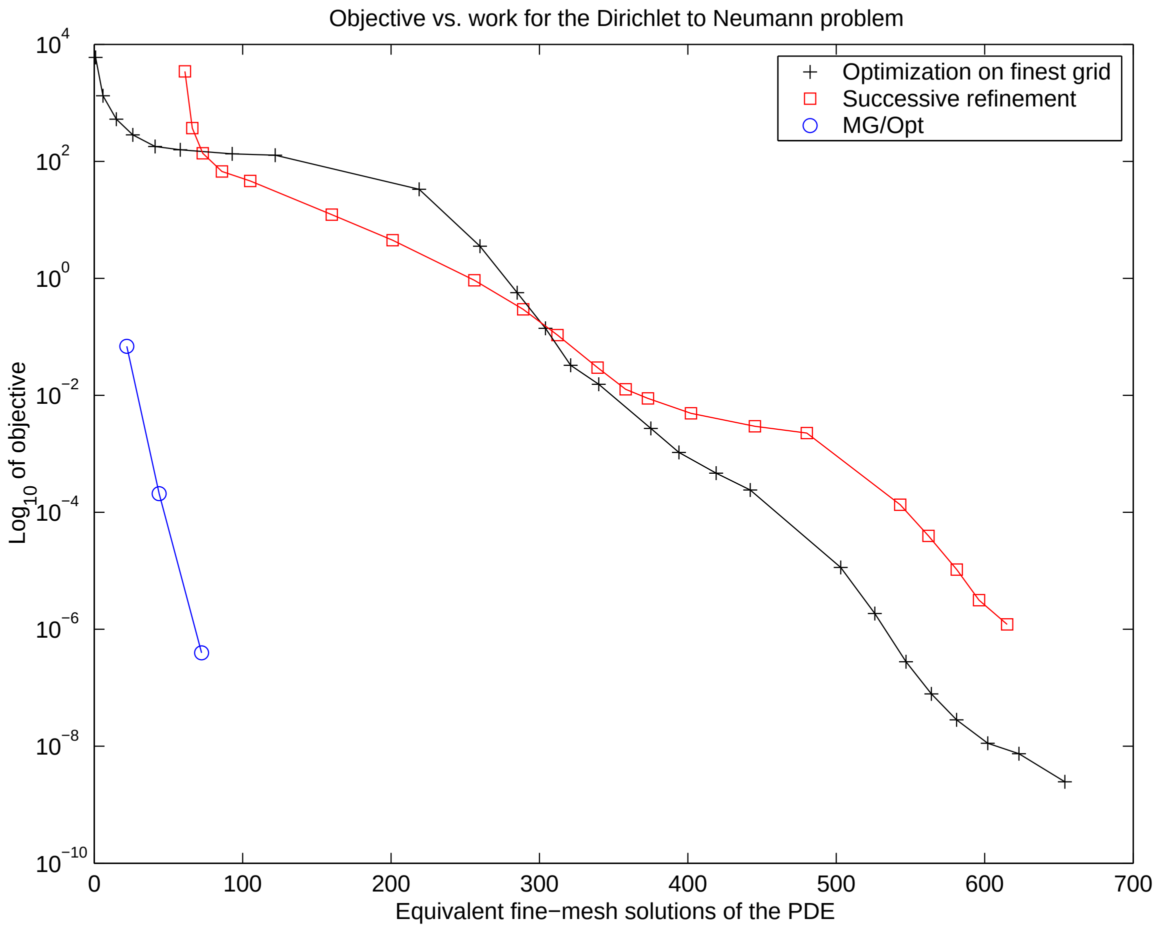
$$\int_0^{\pi} \left[\frac{\partial u}{\partial n}(x_1, 0) - \phi(x_1)\right]^2 dx_1$$

where $S = \{\ (x_1, x_2)\ \mid\ 0 \le x_1 \le \pi,\ 0 \le x_2 \le 1\ \}$

Governing BVP:

$$\begin{aligned} \Delta u &= 0 && \text{on the square } S, \\ u\Big|_{\Gamma} &= a(x_1), && \Gamma = \text{lower edge of } S \\ u\Big|_{\partial S \setminus \Gamma} &= 0 \end{aligned}$$

## Uniform grids (1-d in $a$ and 2-d in $u$): 128, 64, 32, 16



## Model problem: advection

Governing equation: linear advection (hyperbolic):

$$\begin{aligned} u_t + u_x &= 0, \qquad 0 \le t \le T \\ u(x,0) &= a(x) \end{aligned}$$

Objective: minimize

$$F(a) = \frac{1}{2}\int_0^T \int \left| \, u(x,t) - \phi(x,t) \, \right|^2 + \left| \, \partial_x u(x,t) - \partial_x \phi(x,t) \, \right|^2 \; dx \; dt.$$

## The continuous Hessian

The Hessian is given by

$$\nabla^2 F \cdot v = -v''(x) + v(x).$$

This looks ideal for multigrid! BUT. . .

$$\nabla^2 F_h = S_{a,h}^T S_{u,h}^{-T} L_{uu,h} S_{u,h}^{-1} S_{a,h} + S_{a,h}^T S_{u,h}^{-T} L_{ua,h} + L_{au,h} S_{u,h}^{-1} S_{a,h} + L_{aa,h},$$

so it's NOT the case that

$$\nabla^2 F_H = I_H^h \; \nabla^2 F_h \; I_h^H.$$

The situation is more complicated than multigrid applied to equations.

Still, for many problems, we can show that the high-frequency asymptotics are the same for $\nabla^2 F_H$ and $I_H^h \; \nabla^2 F_h \; I_h^H$.

For the model problems, we can compute $\nabla^2 F_h$ directly.

## The discrete Hessian

Forward-time, backwards-space discretization:

$$\frac{u_m^{n+1} - u_m^n}{k} + \frac{u_m^n - u_{m-1}^n}{h} = 0; \quad k = \Delta t, \quad h = \Delta x$$

The discrete Hessian is most simply described in terms of the spatial Fourier transform.

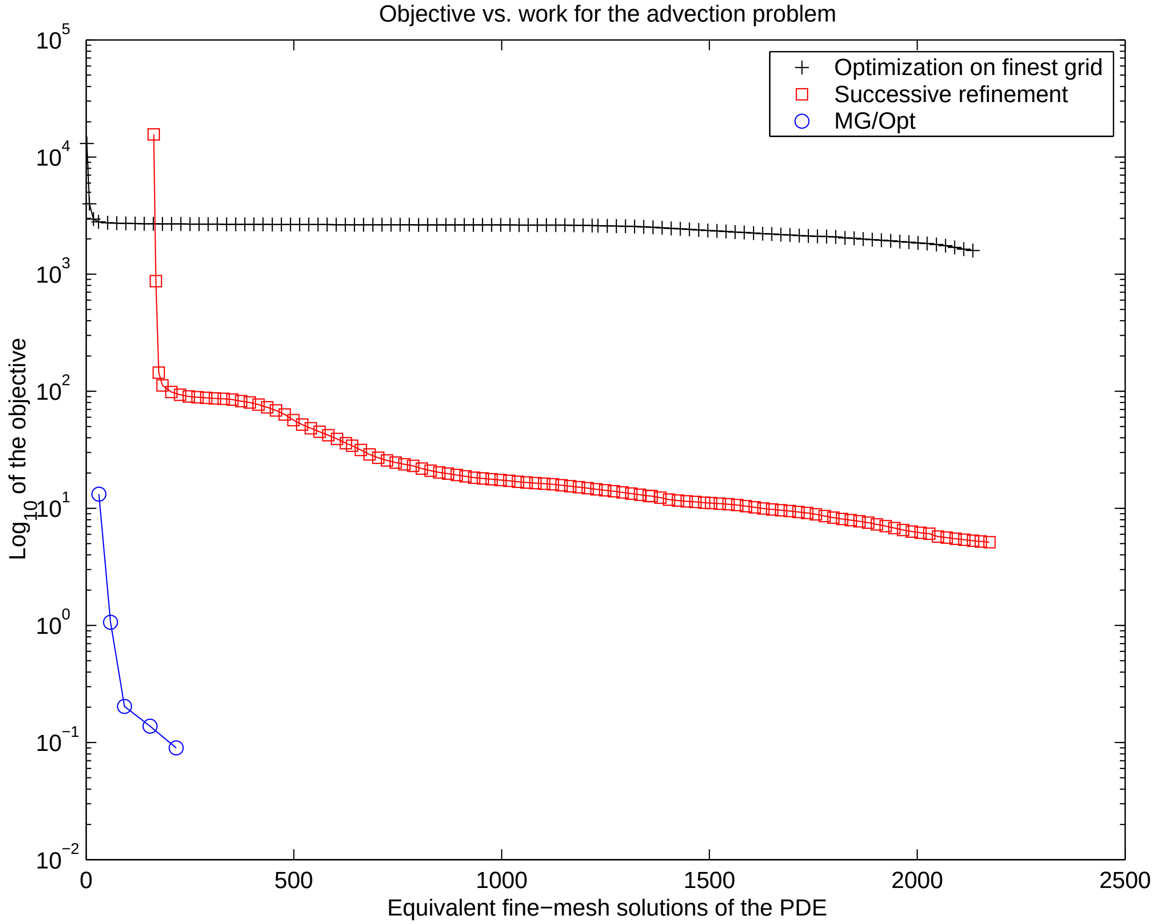
If $\Delta t = \Delta x$ (the stability limit), then

$$(\widehat{\nabla^2 F_h} \cdot v)(\omega) = T\left(1 + \frac{4\sin^2 h\frac{\omega}{2}}{h^2}\right)\hat{v}(\omega) \approx T(1+\omega^2)\hat{v}(\omega).$$

The discrete Hessian looks like an elliptic operator.

Now the analysis begins to resemble classical multigrid.

## Uniform grids (1-d in $a$ and 2-d in $u$): 1024, 512, 256, 128, 64, 32



## Truncated conjugate gradients

In t.c.g., we compute steps by applying c.g. to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\langle s,\ \nabla^2 F\, s\rangle + \langle \nabla F,\ s\rangle \\ \text{subject to} & \qquad \|\ s\ \| \leq \delta. \end{array}$$

We can use the way $\nabla^2 F$ affects low and high frequencies matters to make this more efficient.

## Why does c.g. stall?

Consider unpreconditioned c.g. applied to Poisson's equation, $\Delta u = q$, or, equivalently,

$$\text{minimize}\ \frac{1}{2}\int \nabla u \cdot \nabla u - qu.$$

At iteration $k$, we've minimized over the Krylov subspace

$$\text{“span}\left\{\ q,\ \Delta q,\ \Delta^2 q,\ \Delta^3 q,\ \ldots,\ \Delta^k q\ \right\}\text{”}$$

But the Krylov vectors represent increasingly oscillatory functions, while the solution is *smoother* than $q$ because $\Delta$ is elliptic!

In the discretized problem, c.g. quickly minimizes the quadratic over the span of functions that are increasingly oscillatory *relative to the level of discretization.*

Multigrid switches to coarser levels of discretization to take advantage of this feature.

## Interaction of c.g. and length-scale effects

For a fixed $h$, the discrete Hessian,

$$\nabla^2 F_h(\omega) = T\left(1 + \frac{4\sin^2 h\frac{\omega}{2}}{h^2}\right) \approx 1 + \omega^2,$$

amplifies the upper range of frequencies,

$$|\,\omega\,| \geq \frac{\pi}{2h},$$

more than the lower range,

$$|\,\omega\,| \leq \frac{\pi}{2h}.$$

As in standard MG, in MG/Opt we switch to coarser grids $(h \leftarrow H)$ and apply t.c.g. to knock out the part of the solution that corresponds to the high-frequencies *at that level of discretization.*

## V-cycles

We still need the V-cycle structure of standard MG—we need to do a few iterations of t.c.g. on finer grids from time to time.

The reasons are similar to those in standard multigrid.

The fine-to-coarse grid operators $I_H^h$ are not exact low-pass filters. Since we do not solve the problem exactly on the finer grids, aliasing may occur when we assemble a coarser grid problem.

Conversely, errors can arise when the coarse-grid solutions are injected into the finer grids.

## Is the ellipticity a fluke? Is it expected?

Or is it Egorov's theorem?

Recall the structure of the (reduced) Hessian:

$$\nabla^2 F = S_a^T S_u^{-T} L_{uu} S_u^{-1} S_a + S_a^T S_u^{-T} L_{ua} + L_{au} S_u^{-1} S_a + L_{aa},$$

where $L(a, u) = f(a, u) + \langle \lambda,\ S(a, u) \rangle$.

In this problems (and many realistic problems) we have

$$\nabla^2 F = S_a^T S_u^{-T} L_{uu} S_u^{-1} S_a.$$

The linearized solution operator $S_u^{-1}$ enters via conjugation.

The Hessian frequently turns out to be an elliptic $\Psi$DO, and there are only limited frequency interactions.

Fourier analysis is less suitable for the more general setting.

# Model Problem: Dirichlet to Neumann map (again)

Minimize

$$\int_0^{\pi} \left[ \frac{\partial u}{\partial n}(x_1, 0) - \phi(x_1) \right]^2 dx_1$$

where $S = \{\ (x_1, x_2)\ \mid\ 0 \le x_1 \le \pi,\ 0 \le x_2 \le 1\ \}$

Governing BVP:

$$\Delta u = 0 \qquad \text{on the square } S,$$

$$u\Big|_{\Gamma} = a(x_1), \quad \Gamma = \text{lower edge of } S$$

$$u\Big|_{\partial S \setminus \Gamma} = 0$$

## The analytical Hessian

If

$$v = \sum_{k=1}^{\infty} v_k \sin k x_1,$$

then

$$\nabla^2 F(a) \cdot v \quad = \quad \sum_{k=1}^{\infty} (k^2 \coth^2 k) \; v_k \sin k x_1 \quad \approx \quad -\frac{d^2 v}{d x_1^2},$$

and the Hessian is an elliptic operator.

We would expect multigrid to do well (and CG to do poorly).

# The discrete Hessian

Standard five-point finite-difference scheme:

$$\frac{-u_{m+1,n} + 2u_{m,n} - u_{m-1,n}}{h^2} + \frac{-u_{m,n+1} + 2u_{m,n} - u_{m,n-1}}{h^2} = 0.$$

Grid size $h$ in both the $x_1$ and $x_2$ directions with $h = \pi/N$.

If

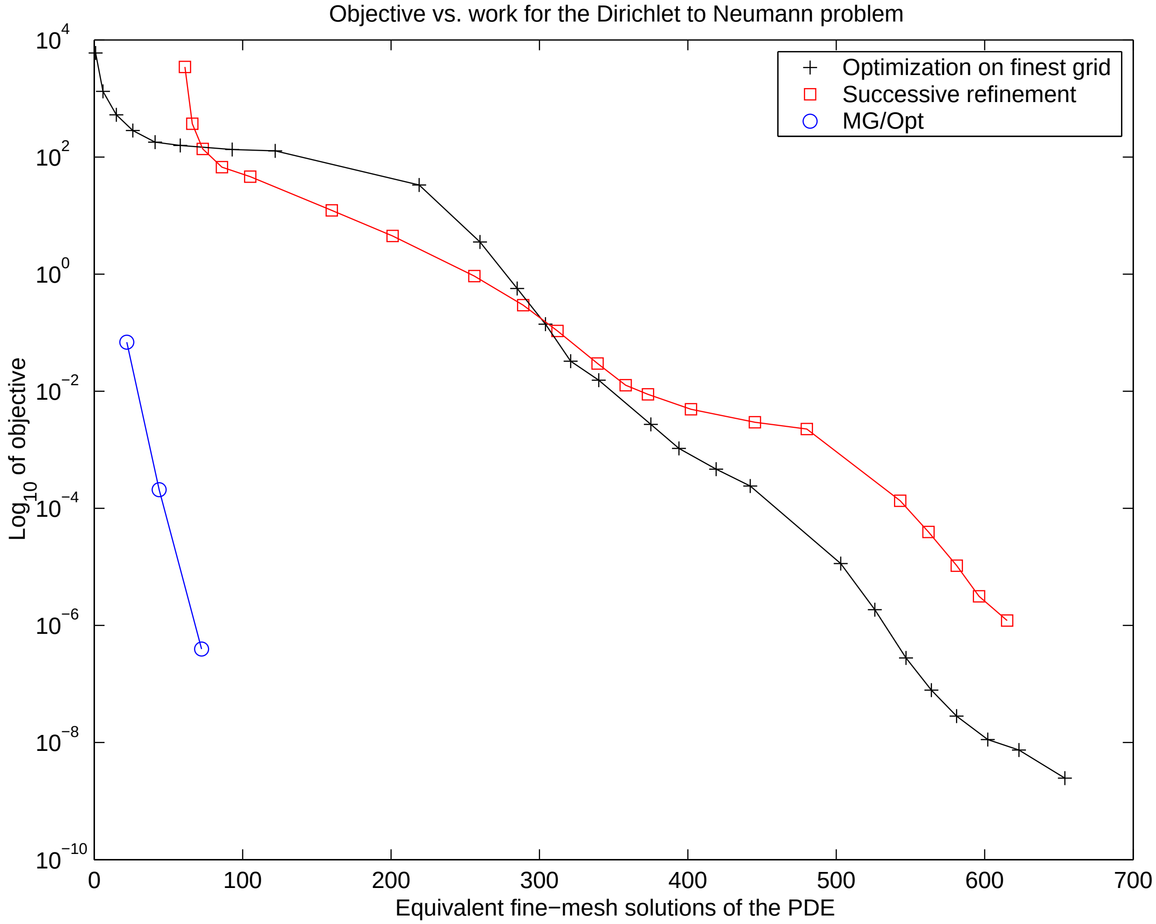
$$v = \sum_{k=1}^{N} v_k \sin kmh,$$

then

$$(Hv)_m = \sum_{k=1}^{N} \sigma_k^2 \, v_k \sin kmh$$

where $\sigma_k^2$ still grows roughly like $k^2$.

## Uniform grids (1-d in $a$ and 2-d in $u$): 128, 64, 32, 16



# Summary

- Multigrid is applicable to optimization of systems governed by differential equation constraints
- Can be successful even if the underlying p.d.e. are not elliptic
- Approach separates model, discretization, and optimization
- Structural features of the reduced Hessian lead us to believe multigrid will be widely applicable

## Interpreting the reduced Hessian

The identity

$$\nabla^2 F(a) = W^T \left( \nabla^2_{(a,u)} L((a, u(a); \lambda) \right) W.$$

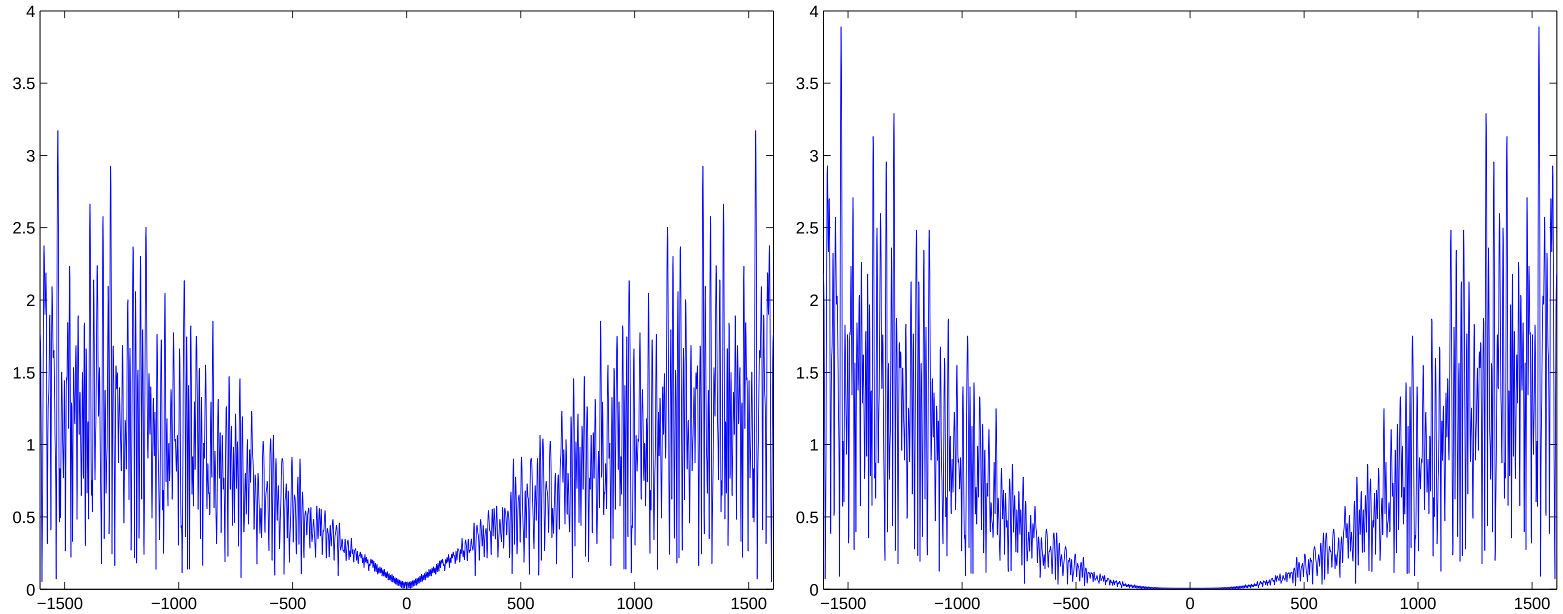
means

$$\nabla^2 F(a)[\eta_1, \eta_2] = \nabla^2_{(a,u)} L((a, u(a); \lambda)[W\eta_1, W\eta_2].$$

*Back to the reduced Hessian*

# Successive Krylov vectors for the advection problem

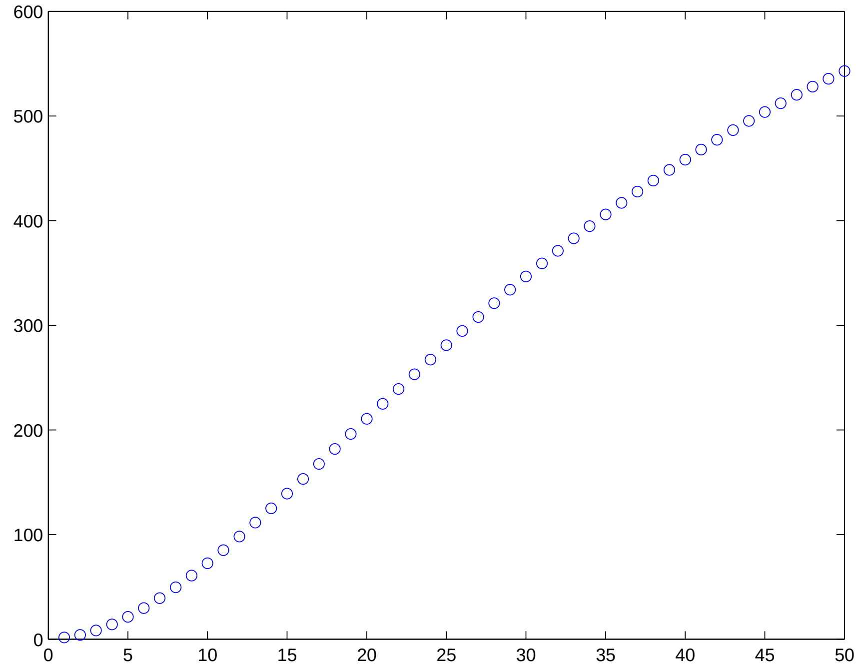
Plot of the magnitudes of the FFTs




*Back to the advection Hessian*

## The discrete Hessian for the Dirichlet to Neumann map

$\sigma_k^2$ versus wavenumber for $h = 0.01$




*Back to the Dirichlet to Neumann map*